Research Brief: Large Language Models for Code Generation

Executive Summary

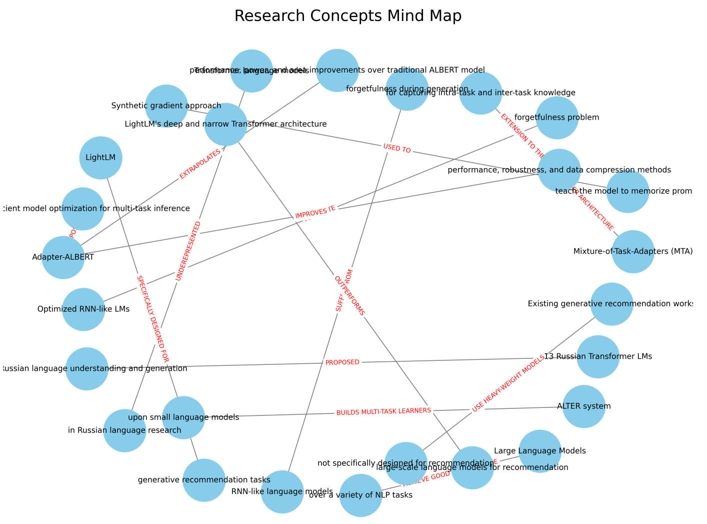
Title: Innovative Language Model Approaches for Efficient Multi-Task Inference

Overarching Trends:
1. Development and optimization of language models (LMs) to address various Natural Language Processing (NLP) tasks simultaneously, reducing computational costs by using small LMs and efficient model architectures.
2. Focus on addressing specific challenges in NLP research, such as prompt forgetting, the lack of suitable models for generative recommendation tasks, and multi-task inference on resource-constrained edge devices.

Common Methodologies:
1. Utilization of Transformer architectures for LMs.
2. Implementation of adapter modules to enhance model performance and flexibility.
3. Two-stage training methods to optimize collaboration between adapters.
4. Synthetic gradient approach to memorize prompts during the generation process in LMs.
5. Indexing methods (Spectral Collaborative Indexing and Graph Collaborative Indexing) to improve model performance for generative recommendation tasks.
6. Experimental evaluation using various datasets and benchmarks.

Potential Resear
1. Exploring appl … and English, such as Arabic, (
2. Investigating t … generative recommendation
3. Examining the … der range of NLP tasks and real-w
4. Further optimi … ce-constrained edge devices to

Concept Mind Map


Research Concepts Mind Map
performance, power, and cost improvements over traditional ALBERT model
Transformer language models
Synthetic gradient approach
LightLM's deep and narrow Transformer architecture
forgetfulness during generation
for capturing intra-task and inter-task knowledge
forgetfulness problem
LightLM
USED TO
EXTRAPOLATES
EXTENSION TO THE
performance, robustness, and data compression methods
teach the model to memorize prom
ARCHITECTURE
cient model optimization for multi-task inference
IMPROVES TE
Adapter-ALBERT
Mixture-of-Task-Adapters (MTA)
UNDEREPRESENTED
SPECIFICALLY DESIGNED FOR
OUTPERFORMS
SUFF ROM
Optimized RNN-like LMs
Existing generative recommendation works
ussian language understanding and generation
PROPOSED
13 Russian Transformer LMs
USE HEAVY-WEIGHT MODELS
upon small language models
BUILDS MULTI-TASK LEARNERS
ALTER system
in Russian language research
Large Language Models
not specifically designed for recommendation
large-scale language models for recommendation
generative recommendation tasks
RNN-like language models
over a variety of NLP tasks
VE GOOD

Detailed Summaries & Sources

ProSG: Using Prompt Synthetic Gradients to Alleviate Prompt Forgetting of RNN-like Language Models

Authors: Haotian Luo, Kunming Wu, Cheng Dai, Sixian Ding, Xinhao Chen

Title: Mitigating Prompt Forgetting in RNN-like Language Models with Synthetic Gradient

Key Findings:

- The study proposes a novel architecture to address prompt forgetting during language model (LM) generation. This issue is particularly problematic when LMs are given complex instructions or prompts.
- By using synthetic gradient, the model is taught to memorize the prompt during the generation process, reducing instances of forgetfulness.

Methodology:

- The proposed method involves deriving states that encode the prompt, transforming them into model parameter modifications via low-rank gradient approximation, effectively hard-coding the prompt into temporary model parameters.
- A dataset is constructed for experimental evaluation.

Main Contribution:

- The paper presents a solution to mitigate prompt forgetting in RNN-like language models, which can significantly improve their performance, especially in scenarios where complex instructions or prompts are involved.

Making Small Language Models Better Multi-task Learners with Mixture-of-Task-Adapters

Authors: Yukang Xie, Chengyu Wang, Junbing Yan, Jiyong Zhou, Feiqi Deng, Jun Huang

Title: ALTER: A System for Multi-Task Learning with Small Language Models

Key Findings & Main Contribution: The research introduces ALTER, a system that utilizes small language models (<1B parameters) to address multiple Natural Language Processing (NLP) tasks simultaneously. ALTER uses the Mixture-of-Task-Adapters (MTA) module as an extension to the transformer architecture, which allows it to capture both intra-task and inter-task knowledge. This approach effectively reduces the computational cost associated with large language models. The two-stage training method proposed further optimizes collaboration between adapters at a minimal computational expense. Experimental results indicate good performance across various NLP tasks, and the study has also resulted in MTA-equipped language models tailored for different domains.

A Family of Pretrained Transformer Language Models for Russian

Authors: Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafuro

LightLM: A Lightweight Deep and Narrow Language Model for Generative Recommendation

Authors: Kai Mei, Yongfeng Zhang

Energy-efficient Task Adaptation for NLP Edge Inference Leveraging Heterogeneous Memory Architect

Authors: Zirui Fu, Aleksandre Avaliani, Marco Donato